稳定的属性，比如昵称、性别、生日、职业、血型、星座、所在地区等；还有在各种行为中产生的文本内容，如在微博上用户根据自己的兴趣爱好发表的博客、转发的帖子、对其他用户帖子的评论等。

虽然目前已有一些结合网络中文本和结构信息进行网络表征的方法，但是针对社交网络中复杂的结构信息和丰富的文本信息，在网络表征的算法中融合何种属性以及如何融合这些属性，仍是未解决的问题。

2　相关研究

在很多网络中文本信息和网络结构是共存的，一些学者开始关注多种信息融合的网络表征。其中有使用神经网络的方法，其典型代表有 Tu 等[11]提出的 CANE 算法，该方法通过加入邻居节点的信息来学习网络中文本的表征，用 CNN 对文本信息进行编码，并引入相互注意力机制，节点邻居不同，学习到的网络表征就不同；Pan 等[12]提出了 TriDNR 算法，该方法利用耦合神经网络体系结构，综合节点结构、节点内容和节点标签（如果可用）来学习最优的节点表示。还有一些学者使用了矩阵建模的方法，把网络结构和文本信息用矩阵统一起来进行网络表征，相关研究有 Huang 等[13]提出的 LANE 算法，该方法将标签信息整合到网络表征中，学习到的网络表征保留了原始网络的拓扑结构和节点属性的相似度；Yang 等[14]结合网络的结构和文本属性，基于矩阵分解提出了 TADW（Text-Associated Deep Walk）方法来学习网络表征。上述方法使用了节点标签信息、节点文本信息、单结构信息，但没有利用多结构信息，而在真实的网络中不仅存在丰富的外部信息，还存在多种网络结构，而多结构中也包含着很多潜在信息。

考虑到网络的多结构性，我们希望网络的表征在考虑外部信息的同时可以结合多种结构信息。经过分析并对比上述算法后发现，TADW 方法对网络表征中多结构和文本信息的融合有一定的启发作用。

TADW 方法在关系矩阵分解时加入了文本特征，既考虑了结构，也结合了文本，但针对矩阵分解的特殊性，文献[14]中并没有对 TADW 方法中文本的融合是否最优进行讨论和分析。因此，本文基于 TADW 方法的原理，通过大量实验论证，提出了一种改进的融合文本属性的网络表征方法。同时，考虑到社交网络中多种结构信息和文本信息并存的情况，本文提出了一种融合多结构信息和丰富文本信息的网络表征方法，以提升网络表征效果。

本文第 3 节简单介绍 TADW 方法，并分析验证 TADW 中文本特征位置对网络表征的影响；第 4 节针对社交网络的特殊性，基于 TADW 方法提出一种多结构及文本融合的网络表征方法；第 5 节通过多分类任务的实验验证本文所提网络表征的有效性。

3　问题模型

3.1　网络模型

社交网络中，用户之间由关注、朋友、粉丝等自报道信息构成关系网络，通过相互转发消息、私信、评论、收藏、点赞、点踩等形成交互网络。两种网络结构既有联系又有区别：交互网络由 $G_{\text{trans}}=(V,E,T)$ 表示，其中 V 是交互网络顶点的集合，$v_i(v_i\in V)$ 表示用户，E 是边的集合，边 $e=(v_i,v_j)\in E$，v_i 和 v_j 表示发生交互的一对用户，T 是用户文本信息的集合；关系网络由 $G_r=(V',E')$ 表示，其中 V' 表示关系网络中用户的集合，$v_i'\in V'$，E' 表示存在关系的用户对的集合，$e'=(v_i',v_j')\in E'$，这里 $V=V'$，即不同网络结构中的用户是相同的。

3.2　TADW 简介

TADW 方法[14]由清华大学的刘知远团队所提，该方法是结合了网络结构和文本进行网络表征的典型方法。网络由 $G=(V,E,T)$ 表示，其中 V 是顶点集合，E 是网络的边集合，T 是所有顶点的文本特征。文献[14]证明了 DeepWalk 等价于矩阵分解 $M=W^{\mathrm{T}}H$，因此本文使用归纳矩阵分解[15]的方法进行矩阵分解，在矩阵分解的过程中加入文本特征来学习网络表征。

TADW 的模型如图 1 所示，其中 $M\in R^{|V|\times|V|}$，M_{ij} 表示 G 中顶点 v_i 随机游走到顶点 v_j 的平均概率的对数，根据 DeepWalk 原理得到。$W\in R^{k\times|V|}$，k 是顶点表征的维度，$|V|$ 是 G 中的顶点个数。$T\in R^{f_t\times|V|}$，T 中每一列对应一个顶点的文本特征表示，f_t 是每个顶点文本特征表示的维度。$H\in R^{k\times f_t}$，这里 $k\ll|V|$。TADW 方法中的顶点表征为 $[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$（$\otimes$ 表示直接连接两个矩阵中的各个分量，上标 T 表示矩阵转置，$*$ 表示矩阵之间的乘积），最终网络中顶点的表征维度为 $2k$。

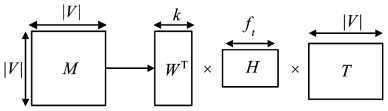

图 1　TADW 模型

Fig. 1　TADW model

归纳矩阵分解即最小化公式（1），对矩阵 M 分解得到 W 和 H，使得 $M\approx X^{\mathrm{T}}W^{\mathrm{T}}HY$。在文献[15]中，$M$ 表示两种属性的关系的观察矩阵，X 表示一种属性的特征矩阵，Y 表示另一种属性的特征矩阵，学习到的 $W^{\mathrm{T}}H$ 是两种属性潜在的关系矩阵。

$$\min_{W,H}\sum_{(i,j)\in\Omega}(M_{ij}-(X^{\mathrm{T}}W^{\mathrm{T}}HY)_{ij})^2+\frac{\lambda}{2}(\|W\|_F^2+\|H\|_F^2) \tag{1}$$

从式（1）可知归纳矩阵分解需要两个特征，为了方便，分别称其为 X 特征和 Y 特征。TADW 方法中只用到了文本特征 T，文献[14]用单位矩阵 E 表示 X 特征，用文本特征矩阵 T 表示 Y 特征。由此，TADW 方法的优化目标变为式（2），矩阵分解的结果为 $M\approx EW^{\mathrm{T}}HT$。

$$\min_{W,H}\sum_{(i,j)\in\Omega}(M_{ij}-(W^{\mathrm{T}}HT)_{ij})^2+\frac{\lambda}{2}(\|W\|_F^2+\|H\|_F^2) \tag{2}$$

3.3　文本特征位置对网络表征的影响

本文首先对融合文本和关系结构的不同网络表征方法的效果进行分析。根据式（1）和 TADW 方法的思想，在等价的矩阵分解模型中，文本特征 T 可以处于 X 或 Y 特征矩阵位置。根据文本特征位置的不同，本文主要分析了 4 种不同表征方法的质量。下面对 4 种网络表征方法进行详细介绍。

当文本特征 T 放置在 Y 特征位置时，网络表征方法分为 $[W^T]$ 和 $[W^T \otimes T^T * H^T]$ 两种。

1) W^T：文献[14]对矩阵分解和 DeepWalk 方法的等价性进行了证明，指出由 DeepWalk 方法得到的网络表征为 W^T。在矩阵分解模型中考虑文本特征时，可得到图 1 所示模型，其中 W^T 已包含若干文本特征或由文本特征带来的影响，可以作为表征学习的一种结果。

2) $[W^T \otimes T^T * H^T]$：为 TADW 方法中的网络表征，即将矩阵分解得到的表征结果与文本特征 T 进行连接后作为最终表征。

当文本特征 T 放置在 X 特征位置时，网络表征方法分为 $[T^T * W^T]$ 和 $[T^T * W^T \otimes H]$ 两种。

1) $[T^T * W^T]$：该表征考虑了文本信息，但没有考虑分解式中 H 隐含的信息。

2) $[T^T * W^T \otimes H]$：将文本特征 T 置于 X 特征位置时的网络表征形式。

下面对这 4 种不同表征的效果进行数值分析。

3.3.1 数据准备和实验设置

在 Cora，Citeseer，Wiki 3 个不同来源的数据集上进行实验，以分析在 TADW 模型的基础上所提到的 4 种不同网络表征方法的性能。这 3 个数据集中，Cora 和 Citeseer 描述了论文间的引用关系。其中，Cora 数据集包含 2708 篇关于机器学习的文章，分为 7 种类型，每篇论文可有多个主题词，共有 1433 个主题词，论文之间的引用关系有 5429 条；Citeseer 数据集包含 3312 篇文章，分为 6 种类型，包含 3703 个主题词，文章之间的引用关系有 4732 条；Wiki 数据集中有 2405 篇文档，19 种类型，包含 4973 个主题词，文档之间存在的联系共有 17981 条。

本实验使用的参数与 TADW 中的参数一样，$f_t=200$，顶点表示维度是 $2k$，正则化参数 $\lambda=0.2$，Cora 和 Citeseer 数据集中 $k=80$，Wiki 数据集中 $k=100$。

在实验数据的预处理中，通过对文档主题使用 TF-IDF 算法得到文本特征矩阵 T；同时，考虑到计算性能，使用 SVD 算法对特征矩阵进行降维处理。在后文的实验中，文本特征矩阵 T 均是经过上述处理得到的。

3.3.2 实验结果及分析

在这 3 个数据集上分别使用上述 4 种不同方法得到网络表征，并用这些表征完成多分类任务，最后用分类准确度评价文本位置对网络表征质量的影响。本实验在进行多分类任务时使用的分类器是基于有监督学习的 SVM[16] 分类器。若分类准确度高，则说明网络表征效果较好。

实验结果如表 1－表 3 所列，其中方法②就是 TADW 原始的表征方法。

表 1 Citeseer 数据集上的分类准确度

Table 1 Classification accuracy on Citeseer dataset

Method	Embedding	Train_radio				
		0.1	0.2	0.3	0.4	0.5
①	$[W^T]$	0.6774	0.7089	0.7194	0.7244	0.7264
②	$[W^T \otimes T^T * H^T]$	0.7016	0.7232	0.7311	0.7367	0.7366
③	$[T^T * W^T]$	0.6950	0.7169	0.7222	0.7276	0.7292
④	$[T^T * W \otimes H]$	**0.7102**	**0.7328**	**0.7390**	**0.7441**	**0.7443**

表 2 Wiki 数据集上的分类准确度

Table 2 Classification accuracy on Wiki dataset

Method	Embedding	Train_radio						
		0.03	0.07	0.1	0.2	0.3	0.4	0.5
①	$[W^T]$	0.4251	0.5872	0.6251	0.6756	0.6903	0.7005	0.7089
②	$[W^T \otimes T^T * H^T]$	0.5798	0.6857	0.7167	0.7599	0.7737	0.7830	0.7916
③	$[T^T * W^T]$	0.5822	0.6780	0.7115	0.7478	0.7599	0.7695	0.7776
④	$[T^T * W \otimes H]$	**0.5902**	**0.6948**	**0.7334**	**0.7757**	**0.7941**	**0.8040**	**0.8096**

表 3 Cora 数据集上的分类准确度

Table 3 Classification accuracy on Cora dataset

Method	Embedding	Train_radio				
		0.1	0.2	0.3	0.4	0.5
①	$[W^T]$	0.8113	0.8485	0.8592	0.8595	0.8633
②	$[W^T \otimes T^T * H^T]$	**0.8218**	**0.8500**	**0.8613**	**0.8674**	**0.8696**
③	$[T^T * W^T]$	0.7290	0.7611	0.7703	0.7736	0.7736
④	$[T^T * W \otimes H]$	0.7852	0.8152	0.8244	0.8288	0.8332

表 1－表 3 列出了 3 个数据集上的分类准确度，其表现出来的共同特征为：连接 W^T 和文本特征的网络表征效果普遍高于没有连接文本特征的结果，即 $[W^T \otimes T^T * H^T]$ 的表征效果优于 $[W^T]$，$[T^T * W^T \otimes H]$ 的效果优于 $[T^T * W^T]$ 的效果。这说明文本特征矩阵 T 不仅在求取表征矩阵时有用，还有助于提升最终的网络表征效果，而且分解得到的 H 矩阵中也包含了网络中隐含的信息，因此文本特征矩阵 T 不可忽略。

在 3 个数据集上表现出的不同结果有：在 Cora 数据集上，文本特征 T 置于 Y 的位置时得到的网络表征效果优于其置于 X 的位置的网络表征效果。而在 Citeseer 和 Wiki 数据集上，文本特征 T 置于 X 的位置时得到的网络表征效果 $[T^T * W^T \otimes H]$ 优于其置于 Y 的位置时得到的网络表征效果。这个区别的产生与数据集中的文本表征矩阵 T 有一定关联。在 Citeseer 和 Wiki 数据集上的主题词数远远多于 Cora 数据集上的主题词数，由于 3 个数据集上的文本特征矩阵 T 是根据文档关键词和 TF-IDF 算法得到的，主题数越多，得到的文本特征矩阵就会越稀疏，虽然经过 SVD 算法降维得到的文本特征矩阵的维度相同，但是在主题词较少的情况下，文本的表征效果会更好。

这个规律可以反映出文本特征在分解式中的位置对网络表征的影响，同时有助于找到更合理的网络表征方法。即当网络顶点文本中所涉及的主题词较少时，$[W^T \otimes T^T * H^T]$ 方法表征的效果会比较好；当主题词较多时，$[T^T * W^T \otimes H]$ 方法表征的效果会更好。

4 多结构及文本融合的网络表征

4.1 社交网络的多结构性

网络表征的一般模型中，都假设顶点间的关系是静态的，网络结构是单一的，而社交网络中多种结构关系并存，结构复杂，同时每个用户会产生丰富的文本信息，在用户节点表征中如果能够融合这些复杂的结构特征及文本属性信息，那么节点表征对用户的表示能力将会更强，而用户分析的结果则会更精准。

4.2　多结构及文本融合的网络表征模型

针对社交网络中交互结构和关系结构并存、属性文本和关系结构混杂的情况，基于 TADW 模型，本文进一步提出了一种多结构及文本融合的网络表征模型（Multi-structure，text-associated DeepWalk，MsTADW），如图 2 所示。其中，$M\in R^{|V|\times|V|}$，M_{ij} 表示交互网络 G_{trans} 的顶点 v_i 游走到顶点 v_j 的平均概率的对数；T 为文本特征矩阵，使用 TF-IDF 算法对用户文本内容计算得到，$T\in R^{f_t\times|V|}$，f_t 是用户文本特征表示的维度；$|V|$ 表示 G_{trans} 中顶点的个数；$W\in R^{k\times f_t}$，这里 $k\ll|V|$；$H\in R^{k\times f_r}$；R 表示关系网络的生成矩阵，由对关系网络 $G_r=(V',E')$ 进行随机游走得到，与 M 矩阵的生成算法相同。然后使用 SVD 降维，$R\in R^{f_r\times|V'|}$，f_r 为关系网络中每个用户节点的表征维度。经过学习，最终的网络节点表征为 $[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes R^{\mathrm{T}}*H^{\mathrm{T}}]$，维度为 $2k$。

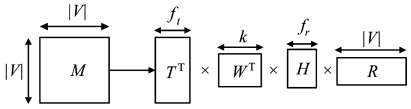

图 2　MsTADW 模型

Fig. 2　MsTADW model

MsTADW 模型中，网络表征学习的优化目标如式（3）所示。使用共轭梯度下降法迭代更新矩阵 W 和矩阵 H，使得 $M\approx T^{\mathrm{T}}W^{\mathrm{T}}HR$。

$$\min_{W,H}\sum_{(i,j)\in\Omega}(M_{ij}-(T^{\mathrm{T}}W^{\mathrm{T}}HR)_{ij})^2+\frac{\lambda}{2}(\|W\|_F^2\|H\|_F^2) \tag{3}$$

MsTADW 针对社交网络的特殊性，结合了网络中的多种信息进行网络节点表征。为验证该方法的有效性，在多个数据集上进行实验，并将其与传统的网络表征方法进行比较分析。

5　实验及结果分析

5.1　实验数据及参数设置

在实验数据集的选择和参数设置上，考虑到第 3 节实验中的 Cora，Wiki，Citeseer 3 个数据集上仅有一种结构信息，即论文引用关系，缺少多种结构信息，本次实验中没有继续使用这 3 个数据集，而使用了采自新浪微博的数据集 Microblog。该数据集采集了 2009 年 8 月 28 日到 2012 年 12 月 26 日新浪微博 1701 个用户的行为信息，包括用户之间互加好友的信息、相互转发微博和评论的记录，以及发表微博的内容、评论的内容等。其中，用户转发记录 90962 条，好友关系 29439 个，用户兴趣分为 10 类，主题词是对用户发表博客、转发博客、评论博客的内容进行分词、去停顿词、去重后得到的，共有 68362 个。

本实验中的交互结构由用户转发微博的行为构成，关系结构由用户互加好友、互相关注的行为构成，其中 $|V|=1701$，$f_r=200$，其余参数与 TADW 的参数一样，$f_t=200$，$k=100$，正则化参数 $\lambda=0.2$，顶点的表征维度为 $2k$。

网络表征的效果仍然采用其在同一分类任务上的分类质量进行评估。实验中采用 SVM 分类器对用户兴趣分类的准确度来验证不同网络表征的效果，将在 Microblog 数据集上使用的训练集比率分别设定为 10%，20%，30%，40%，50%。

5.2　多结构及文本融合对网络表征效果的影响

为了验证本文提出的 MsTADW 模型的有效性，选取了 DeepWalk，TADW，RADW 作为基准比较对象。

DeepWalk 方法是一种基于交互结构的网络表征。根据式（1），DeepWalk 方法中的 X 特征和 Y 特征都是用单位矩阵 E 来表示的，网络表征为 $[W^{\mathrm{T}}]$。

TADW 方法是融合了交互结构和文本信息的网络表征方法，该方法按照矩阵分解时特征矩阵位置的不同有两种表征，分别为 $[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$，$[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$，为了方便，将两种表征方法分别命名为 TADW1，TADW2。

RADW 是本文为了分析两种不同关系结构信息对网络表征的影响，在 TADW 模型中用关系结构特征 R 代替文本特征 T 得到的网络表征方法，其网络表征为 $[W^{\mathrm{T}}\otimes R^{\mathrm{T}}*H^{\mathrm{T}}]$。

第 3 节已验证了在 TADW 方法中文本特征在 X 特征位置和 Y 特征位置时学习到的网络表征的效果是不同的，这里在 Microblog 数据集上也对 $[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$ 和 $[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$ 方法进行了对比实验。

最后，将 MsTADW 与 DeepWalk，RADW 和 TADW 方法进行对比实验，以验证多结构及文本融合的网络表征方法的有效性。实验结果如表 4 所列。

表 4　多结构对网络表征效果的影响

Table 4　Influence of Multi-structures on network representation

Method	Embedding	Train_radio				
		0.1	0.2	0.3	0.4	0.5
DeepWalk	$[W^{\mathrm{T}}]$	0.2348	0.2443	0.2473	0.2469	0.2515
RADW	$[W^{\mathrm{T}}\otimes R^{\mathrm{T}}*H^{\mathrm{T}}]$	0.4100	0.4518	0.4622	0.4640	0.4748
TADW1	$[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$	0.3720	0.4350	0.4413	0.4371	0.4492
TADW2	$[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$	0.3904	0.4624	0.4734	0.4727	0.4959
MsTADW	$[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes R^{\mathrm{T}}*H^{\mathrm{T}}]$	**0.4348**	**0.4666**	**0.4753**	**0.4862**	**0.4922**

表 4 的实验结果表明，在多种训练比率的情况下，RADW 用于分类的准确率比 DeepWalk 的准确率平均高出 20%左右，网络表征效果有了显著的提升。

TADW2 方法的 $[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$ 表征效果比 TADW1 方法的 $[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$ 好，说明在 Microblog 数据集上当文本特征放在 X 特征的位置时，表征效果更好。在该数据集上的主题词有上万个，从而也说明了当主题词较多时，$[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$ 方法的表征效果会更好，同时验证了本文对文本特征位置对网络表征影响分析的正确性。

MsTADW 方法的准确率比 RADW 方法提高了 2%左右，MsTADW 方法比 TADW1 方法中表征为 $[W^{\mathrm{T}}\otimes T^{\mathrm{T}}*H^{\mathrm{T}}]$ 的准确率提高了 5%左右；而在训练比率较低的情况下，其准确度高于表征为 $[T^{\mathrm{T}}*W^{\mathrm{T}}\otimes H]$ 的方法。该结果说明了对于社交网络，文本特征和结构特征同等重要，在训练样本较少的情况下，网络中的各种信息更能相互补充，达到多信息有

（下转第 77 页）